Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Matjaž Balon

Ekstremno naključni kvantilni gozdovi

diplomsko delo
na univerzitetnem študiju

Mentor: prof. dr. Igor Kononenko

Ljubljana 2016

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Ena uspešnejših metod v strojnem učenju so naključni gozdovi. Ekstremno naključni gozdovi dodajajo večjo naključnost, ekstremno naključni kvantilni gozdovi pa omogočajo napovedovanje intervalov pri regresijksih napovedih, kar je koristno pri podatkih s spremenljivo varianco. V okviru diplomske naloge implementirajte različne variante naključnih gozdov in primerjajte njihovo uspešnost na umetnih in realnih podatkovnih množicah tako, da uporabite pokrivno verjetnost, relativni povprečni interval in kombinirano statistiko. Analizirajte tudi vplive različnih parametrov, kot so število dreves, velikost učne množice, težavnost ciljne funkcije, šum v podatkih, na uspešnost napovedi ter na časovno zahtevnost učenja in napovedovanja. Rezultate primerjajte tudi z metodami najbližjih sosedov (NS), stremljenja in maksimalnega verjetja (SMV) ter njune kombinacije NS-SMV, ki jih je v svoji disertaciji razvil Darko Pevec.

Izjava o avtorstvu diplomskega dela

Spodaj podpisani Matjaž Balon sem avtor diplomskega dela z naslovom:

Ekstremno naključni kvantilni gozdovi

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom prof. dr. Igorja Kononeka,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela,
- soglašam z javno objavo elektronske oblike diplomskega dela na svetovnem spletu preko univerzitetnega spletnega arhiva.

V Ljubljani, dne 5. maja 2016 Podpis avtorja:

Za mentorstvo, vodenje in koristne komentarje o delu se zahvaljujem prof. dr. Igorju Kononenku. Zahvala gre tudi as. dr. Darku Pevcu, ki je bil v veliko pomoč z nasveti in napotki. Posebna zahvala pa gre mami, ki mi je stala ob strani vsa leta študija.

Kazalo

Povzetek

Ekstremno naključni kvantilni gozdovi so ansambelska metoda, ki z vnašanjem dodatne naključnosti in kvantilov razširja običajne naključne gozdove. V tem delu smo preverili njeno napovedno točnost z različnimi merami uspešnosti ter hitrost izvajanja. Analizirali smo tudi vplive različnih parametrov in velikosti učne množice na uspešnost delovanja in čas potreben pri izvajanju. Metodo smo primerjali z različnimi merami uspešnosti in časi delovanja še z dvema metodama, ki za napovedovanje uporabljata kvantile in s tremi metodami, ki dajejo napovedne intervale. Ekstremno naključni kvantilni gozdovi so se izkazali kot zelo konkurenčni tako v točnosti napovedovanja po različnih merah kot tudi v hitrosti izvajanja.

Ključne besede: strojno učenje, kvantili, naključni gozdovi, regresija.

Abstract

Extremely randomized quantile forests are an ensemble method which extends ordinary random forests with additional randomness and quantiles. In this work we checked its prediction accuracy with different success measures and execution speed. We also analyzed influences of various parameters and size of learning dataset on prediction performance and time needed for execution. We compared method with different success measures and execution time also with two methods that use quantiles for prediction and three methods that give prediction intervals. Extremely randomized quantile forests have been proved as competitive in terms of prediction strength with different measures and also in speed of execution.

Keywords: machine learning, quantiles, random forest, regression.

Poglavje 1

Uvod

V strojnem učenju [13] poznamo veliko metod za reševanje problemov, tako klasifikacijskih kot tudi regresijskih. Ena bolj znanih in uspešnih metod so naključni gozdovi [4]. Gre za ansambelsko metodo, ki temelji na odločitvenih drevesih in odpravlja njihove pomanjkljivosti. Obstaja precej razširitev te metode. Ekstremno naključni gozdovi so ena izmed njih in vnašajo dodatno naključnost [10]. Ekstremno naključni kvantilni gozdovi pa s kvantili omogočajo napovedovanje intervalov pri regresiji. Slednje je velikokrat zaželeno in pogosto bolj natančno kot napovedovanje točnih vrednosti pri običajnih regresijskih metodah. Pri tem so kvantili zelo informativni predvsem na heteroskedastičnih podatkih. Heteroskedastičnost pomeni, da se razpršenost med opazovanji spreminja. Varianca torej lahko z opazovanji narašča ali pada oziroma se kako drugače spreminja.

V delu smo naredili pregled metod ter preverili njihovo delovanje na umetnih in realnih podatkovnih množicah. Analizirali smo vpliv različnih parametrov posameznih metod, velikost učne množice, težavnost ciljne funkcije in šum v podatkih na uspešnost napovedi in čas delovanja za učenje in napovedovanje. Rezultate analize smo primerjali še z metodami najbližjih sosedov, stremljenja in maksimalnega verjetja ter njune kombinacije.

V naslednjem poglavju je pregled metod iz katerih izhajajo ekstremno naključni kvantilni gozdovi, ter opis metod s katerimi jih primerjamo. V tre-

tjem poglavju predstavimo metodologijo primerjave metod. Četrto poglavje predstavi rezultate in analizo primerjav. V petem poglavju pa so povzete ugotovitve in zaključki.

Poglavje 2

Pregled metod

Metode modeliranja podatkov običajno delimo na nadzorovane in nenadzorovane. Pri nenadzorovanih skuša algoritem izluščiti število skupin in njihove skupne lastnosti iz vhodnih podatkov. Mi pa smo se osredotočili na nadzorovano učenje, kjer za učenje uporabljamo vhodne in izhodne podatke, na podlagi katerih lahko po učenju model daje napovedi za vhodne podatke pri katerih so izhodni podatki neznani.

2.1 Naključni gozdovi

Naključne gozdove (angleško *Random forest*) je leta 2001 predlagal Breiman [4]. Gre za ansambelsko metodo, ki združuje odločitvena drevesa (angleško *Decision trees*). Ker ta veljajo za nestabilen algoritem [7], jih lahko z vnosom naključnosti uspešno združujemo v ansambelsko metodo [6].

Naključni gozdovi gradijo drevesa tako, da za vsako drevo iz osnovne učne množice naredijo novo učno množico s pomočjo vzorčenja z zamenjavo (angleško *Bagging*) [3]. Pri tem postopku iz osnovne učne množice naključno izberemo enako število primerov, kot jih je v osnovni učni množici s tem, da se primeri lahko ponavljajo. Tako dobimo učne množice, ki imajo v povprečju 63,2% različnih primerov, ostali pa so ponovljeni. Na takšnih učnih množicah se nato gradijo drevesa, ki pa za razliko od običajnih odločitvenih dreves v

vozliščih ne izbirajo najboljšega atributa med vsemi možnimi, ampak med določenim številom naključno izbranih. Tako dobimo dovolj različna drevesa, da ansambelska metoda deluje dobro.

Drevesa gradimo do največje globine in jih ne režemo. Vsako tako drevo s svojo napovedjo nato glasuje in z večinskim glasovanjem dobimo končno napoved. S tem se izognemo pristranskosti, ki se lahko pojavi pri odločitvenih drevesih, hkrati pa povečamo natančnost napovedi. Algoritem je tudi precej robusten zaradi izbire učne množice za posamezno drevo in s tem precej odporen na šum v podatkih.

Naključni gozdovi imajo dva parametra. To je število dreves, ki jih zgradimo in število atributov, med katerimi se v vsakem vozlišču izbira najboljšega. Večje kot je število dreves, prej se klasifikacijska napaka ustali. V praksi pa se izkaže, da precej dobre rezultate dobimo že s sto drevesi. Za število atributov Breiman predlaga $\sqrt{n}$, pri čemer je n število vseh atributov v učni množici.

Prednost naključnih gozdov je tudi to, da osnovne učne množice ni potrebno dodatno deliti na testno množico, ker imamo pri vsakem drevesu v povprečju približno tretjino vseh primerov, ki niso bili uporabljeni v učni množici. To so tako imenovani *out-of-bag* primeri. Za vsako drevo naredimo napovedi za vse njegove *out-of-bag* primere in izračunamo razmerje med napačno napovedanimi in vsemi primeri. Povprečje teh razmerij nam da dober približek napake.

2.2 Ekstremno naključna drevesa

Podobno kot naključni gozdovi tudi ekstremno naključna drevesa (angleško *Extremely randomized trees*) [10] temeljijo na odločitvenih drevesih. Kot sledi že iz imena, vnašajo v drevesa še več naključnosti, s katero želijo zmanjšati pristranskost in varianco posameznih dreves, ter tako povečati skupno točnost napovedi. Za razliko od naključnih gozdov na osnovni učni množici ne delajo vzorčenja z zamenjavo, ampak uporabijo celo učno množico. S tem so

računsko bolj učinkoviti, sploh pri velikih učnih množicah.

Ekstremno naključna drevesa gradijo neporezana običajna odločitvena drevesa, le da v vsakem vozlišču naključno izberejo določeno število atributov in točke delitve, ter nato izmed njih izberejo najboljši atribut za dano vozlišče. Končno napoved pa dobimo enako kot pri naključnih gozdovih z večinskim glasovanjem vseh dreves pri klasifikaciji oziroma s povprečjem vseh napovedi pri regresiji.

Za delovanje potrebujejo tri parametre: K, število atributov, ki jih naključno izberejo v vsakem vozlišču, n_{min}, minimalno število primerov za deljenje vozlišča in M, število dreves v ansamblu. Podobno kot pri naključnih gozdovih tudi tu večje število dreves vodi k zmanjšanju napovedne napake. Prav tako v praksi običajno zadošča 100 dreves. Parameter K je med 1 in n, kjer je n število vseh atributov. Manjši kot je K, večja je naključnost dreves in manjša je njihova odvisnost od izhodnih vrednosti učne množice. Pri $K = 1$ je tako ekstremen primer, kjer je izbira atributa popolnoma neodvisna od izhodnih vrednosti učne množice. Drug ekstrem pa je pri $K = n$, kjer ni več naključnosti pri izbiri atributa, ampak samo še pri izbiri točke deljenja. Za klasifikacijske probleme se privzeto uporablja $K = \sqrt{n}$, za regresijske pa $K = n$. Privzeta vrednost minimalnega števila primerov v vozlišču pri klasifikaciji je $n_{min} = 2$, pri regresiji pa $n_{min} = 5$. Večje kot je minimalno število primerov v vozlišču, manjša so drevesa, vendar imajo višjo pristranskost in manjšo varianco [10].

2.3 Kvantilna regresija

Pri običajni regresiji, na primer z metodo najmanjših kvadratov [16], dobimo zgolj točkovne cenilke, ki običajno ne opisujejo najbolje distribucije odvisne spremenljivke. Kvantilna regresija (angleško *Quantile Regression*) [12] pa nam da intervalno cenilko, s pomočjo katere lahko dobimo bistveno boljšo sliko pogojne distribucije. Prav tako je bolj robustna in omogoča odkrivanje osamelcev [23].

2.3.1 Kvantili

Kvantili so točke, ki delijo urejeno množico števil na enako velike dele. Najbolj znani kvantili so: mediana (0.5 kvantil), deli množico na dva dela, tercili, delijo na tri dele, kvartili, delijo na štiri dele, decili, delijo na deset delov in percentili, ki razdelijo množico na sto enakih delov. V splošnem pa so kvantili med 0 in 1, ter delijo množico na poljubno število enakih delov.

Za zvezno porazdelitveno funkcijo je α-kvantil $Q_\alpha(x)$ definiran tako, da je verjetnost, da je Y manjši od $Q_\alpha(x)$ za dani $X = x$ enaka α. V splošnem

$$Q_\alpha(x) = inf\{y : F(y|X = x) >= \alpha\} \tag{2.1}$$

S pari kvantilov pa dobimo napovedne intervale. 90% napovedni interval je tako predstavljen z $I(x) = [Q_{0.05}(x), Q_{0.95}(x)]$

Zgornje je povzeto po [15].

2.4 Kvantilni regresijski gozdovi

Tako kot pri kvantilni regresiji tudi kvantilni regresijski gozdovi (angleško *Quantile Regression Forests*) [15] namesto točkovnih cenilk dajejo intervalne cenilke oziroma pogojne kvantile. Ti z veliko verjetnostjo napovedujejo, kje se bo nahajala vrednost novega opazovanega primera. Velikost napovednega intervala pa je lahko zelo različna. Seveda je pri večjih intervalih napoved slabša in obratno, pri manjših bolj zanesljiva.

Tudi kvantilni regresijski gozdovi temeljijo na naključnih gozdovih. Najpomembnejša razlika med njimi pa je, da naključni gozdovi v listih dreves hranijo samo povprečne vrednosti vseh učnih primerov, kvantilni regresijski gozdovi pa imajo v listih vse opazovane primere. Tako se ohranijo vse informacije o pogojni distribuciji, ki se pri naključnih gozdovih izgubijo.

Algoritem zgradi K dreves kot pri naključnih gozdovih, vendar v vsakem listu zabeleži vsa opazovanja, ne samo povprečja. Vsak nov primer spusti čez vsa drevesa in izračuna uteži. Uteži so relativne frekvence števila primerov

v listih. Vsota vseh uteži posameznega drevesa je ena. Nato izračuna povprečne uteži preko vseh dreves ($w_i(x)$). S temi izračuna oceno distribucijske funkcije za vse y v učni množici. Oceno distribucijske funkcije zapišemo kot

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x) 1_{\{Y_i \leq y\}} \tag{2.2}$$

Pri tem so $1_{\{Y_i \leq y\}}$ indikatorske spremenljivke, ki imajo vrednost 1, ko je $Y_i \leq y$ in 0 sicer.

Pogojne kvantile Q_α pa dobimo z enačbo

$$Q_\alpha(x) = \inf\{y : \hat{F}(y|X = x) \geq \alpha\} \tag{2.3}$$

Za 1 - p napovedne intervale ja nato spodnja meja $Q_{\frac{p}{2}}$ in zgornja $Q_{1-\frac{p}{2}}$.

2.5 Ekstremno naključni kvantilni gozdovi

Ekstremno naključni kvantilni gozdovi so kombinacija ekstremno naključnih dreves in kvantilnih regresijskih gozdov. Izkoriščajo kvantile za napovedovanje intervalov pri regresiji, ob tem pa s povečano naključnostjo stremijo k večji napovedni točnosti in hitrejšem delovanju.

Delujejo podobno kot kvantilni regresijski gozdovi, le da pri grajenju dreves ne delajo vzorčenja z zamenjavo za izdelavo učne množice, ampak uporabijo vse primere pri vseh drevesih, tako kot to počnejo ekstremno naključna drevesa. In enako kot pri slednjih tudi ekstremno naključni kvantilni gozdovi naključno izbirajo atribute in točke delitve pri gradnji posameznih dreves, ter s tem vnašajo večjo naključnost.

Ekstremno naključni kvantilni gozdovi za delovanje potrebujejo enake parametre kot ekstremno naključna drevesa (K, n_{min} in M) in dodaten parameter α, s katerim povemo, za kateri kvantil nas zanimajo napovedi. Privzete vrednosti za parametre K, n_{min} in M so enake kot pri ekstremno naključnih drevesih, za vrednost α pa se najpogosteje uporabljajo vrednosti 0.01, 0.05, 0.25, 0.5, 0,75, 0.95 in 0.99.

2.6 Najbližji sosedi

Metoda najbližjih sosedov (angleško Nearest neighbors) [1] je ena najpreprostejših napovednih metod. Učni del algoritma si zgolj zapomni vse primere. Pri napovedi odvisne spremenljivke za nove primere pa se poišče vnaprej izbrano število najbolj podobnih primerov v atributnem prostoru. Ti primeri nato večinsko klasificirajo nove primere oziroma s povprečjem določijo vrednost odvisne spremenljivke pri regresijskih problemih. Pri tem je potrebno definirati razdalje za diskretne atribute in normalizirati vrednosti zveznih atributov.

Najbližji sosedi za delovanje potrebujejo samo en parameter (k). Ta določa število najbolj podobnih primerov, ki so potrebni za napovedovanje vrednosti odvisne spremenljivke novih primerov. S primerno izbrano vrednostjo tega parametra se lahko bistveno zmanjša vpliv šuma. V praksi je pogosto najbolj primerna vrednost $k = 11$.

Za napovedovanje intervalov s to metodo moramo najprej izračunati predznačene residuale za vse učne primere. Povprečna vrednost residualov $\vec{r}$ nam koristi za popravljanje sredine napovednega intervala in odpravljanje pristranskosti. Varianca residualov $\hat{\sigma}^2(\vec{x})$ definira širino napovednega intervala, ki ga lahko zapišemo z

$$\hat{y}(\vec{x}) + \vec{r} \pm z_{\frac{\alpha}{2}} \hat{\sigma}^2(\vec{x}), \tag{2.4}$$

kjer je z funkcija kumulativne normalne distribucije.

Zgornje je povzeto po [18].

2.7 Stremljenje in maksimalno verjetje (SMV)

Z metodami stremljenja (angleško *Bootstraping*) [14] lahko dobimo dobro oceno variance negotovosti modela $\hat{\sigma}^2_m(\vec{x})$. Metode maksimalnega verjetja (angleško *Maximum likelihood estimation*) [22] pa dajo dobro oceno variance šuma podatkov $\hat{\sigma}^2_p(\vec{x})$.

Na učnih podatkih izvedemo bagging s pomočjo danega učnega algoritma in dobimo napovedi $\hat{y}_{bag}(\vec{x})$. Varianca teh napovedi predstavlja varianco negotovosti modela $\hat{\sigma}^2_m(\vec{x})$. Ker za intervale zaupanja predpostavljamo normalno distribucijo napovedi in pri nesimetrični distribuciji residualov to ne drži, se tam pojavijo širši intervali.

Iz napovedi modela *bagging* uporabimo primere, ki niso bili uporabljeni v učni množici in izračunamo residuale (razlike med pravimi in napovedanimi vrednostmi). V teh residualih so informacije o varianci šuma podatkov, ki jo dobimo z ocenjevanjem maksimalnega verjetja. Za slednje lahko uporabimo mrežo radialnih baznih funkcij.

Napovedni intervali SMV so tako definirani kot

$$\hat{y}(\vec{x}) \pm z_{\frac{\alpha}{2}}\hat{\sigma}^2(\vec{x}) = \hat{y}(\vec{x}) \pm z_{\frac{\alpha}{2}}(\hat{\sigma}^2_m(\vec{x}) + \hat{\sigma}^2_p(\vec{x})) \tag{2.5}$$

Pri tem sta $\hat{\sigma}^2(\vec{x}) = (\hat{\sigma}^2_m(\vec{x}) + \hat{\sigma}^2_p(\vec{x}))$ in $z_{\frac{\alpha}{2}}$ standardna vrednost za izbrano α.

Zgornje je povzeto po [18].

2.8 Kombinirana metoda NS-SMV

Ker metoda najbližjih sosedov tvori bolj optimalne napovedne intervale, metode stremljenja in maksimalnega verjetja pa bolj pravilne, je smiselno njuno združevanje. Pri tem za bolj optimalne napovedne intervale veljajo tisti, ki so ožji. Za bolj pravilne napovedne intervale pa veljajo tisti, ki se čimbolj približajo zastavljeni vrednosti PVNI.

Za izračun napovednih intervalov kombinirane metode NS-SMV enostavno izračunamo povprečje spodnjih napovednih mej obeh metod za spodnjo mejo in za zgornjo mejo povprečje zgornjih napovednih mej obeh metod. To lahko zapišemo kot

$$NS\text{-}SMV\perp = \frac{NS\perp + SMV\perp}{2} \text{ in } NS\text{-}SMV\top = \frac{NS\top + SMV\top}{2} \tag{2.6}$$

Pri tem je NS-SMV$\perp$ spodnja meja in NS-SMV$\top$ zgornja meja napovednega intervala.

Zgornje je povzeto po [18].

Poglavje 3

Metodologija primerjave

Primerjamo uspešnost kvantilne regresije (KR), kvantilnih regresijskih gozdov (KRG), ekstremno naključnih kvantilnih gozdov (ENKG), najbližjih sosedov (NS), stremljenja in maksimalnega verjetja (SMV) ter njune kombinacije (NS-SMV). Pri naključnih kvantilnih gozdovih in ekstremno naključnih kvantilnih gozdovih preizkusimo tudi različne možnosti za izbiro števila atributov oziroma točk rezanja. Vrednosti K so tako 1, $\sqrt{n}$ in n. Pri tem za vrednosti $K = 1$ in $K = n$ to označimo nadpisano pri posamezni metodi (na primer: KRG^1). Privzeta vrednost za obe metodi pa je $K = \sqrt{n}$. Število dreves, ki jih gradita metodi, je nastavljeno na 100, število primerov v listih pa je 5. Napovedi so narejene za 0.01, 0.05, 0.25, 0.5, 0.75, 0.95 in 0.99 kvantile.

3.1 Množice podatkov

Za testiranje smo uporabili 9 umetnih in 5 realnih podatkovnih množic. Osnovne umetne množice imajo 200 primerov in imajo pet neodvisnih spremenljivk. Ena predstavlja linearni problem (LIN) brez šuma. Štiri množice predstavljajo kvadratne probleme. Prva (SQR) je brez šuma, druga (SRN) z dodanim manjšim deležem naključnega šuma, tretja (SGN) s približno 10% Gaussovega šuma in četrta z odvečno spremenljivko. Še dve množici sta s si-

nusnim problemom, kjer je ena (SNS) brez šuma, druga (SLN) pa s približno 10% linearnega šuma. Ena množica predstavlja kompleksen problem (CPX), ki vsebuje tako kvadratno funkcijo kot tudi trigonometrične, dodanega pa ima tudi približno 10% Gaussovega šuma. Zadnja množica (RND) pa ima naključno generirano odvisno spremenljivko.

Realne množice so iz paketa *R* [20] *airquality* (6 spremenljivk, 111 primerov) in *swiss* (6 spremenljivk, 47 primerov). Iz paketa *gss* [11] sta *NO2* (6 spremenljivk, 500 primerov) in *ozone* (10 spremenljivk, 330 primerov). Iz paketa *MASS* [24] pa je še množica *Boston* (14 spremenljivk, 506 primerov). Najmanjša množica podatkov vsebuje 47 primerov, največja 506, v povprečju pa 307 primerov. Iz vseh podatkovnih množic so odstranjeni primeri z neznanimi vrednostmi. Povzete so v tabeli 3.1.

množica	število spremenljivk	število primerov
LIN	6	200
SQR	6	200
SRN	6	200
SGN	6	200
SRV	6	200
SNS	6	200
SLN	6	200
CPX	6	200
RND	6	200
airquality	6	111
Boston	14	506
NO2	6	500
ozone	10	330
swiss	6	47

Tabela 3.1: Umetne in realne podatkovne množice

3.2 Mere uspešnosti

3.2.1 Standardna napaka

S standardno napako (angleško *Standard error*) [2] ocenimo delovanje algoritmov pri kvantilu $\alpha = 0.5$, torej za mediano. Standardna napaka nam pove, kako variirajo posamezne napovedi okrog dejanskih vrednosti odvisne spremenljivke. Definirana je kot

$$S_\epsilon = \sqrt{\frac{\sum (y - y')^2}{n}} \tag{3.1}$$

Pri tem je n število primerov, y dejanska vrednost, y' pa napovedana vrednost odvisne spremenljivke.

3.2.2 Normalizirana standardna napaka

Standardno napako normaliziramo s standardno napako metode, ki vedno napove povprečno vrednost odvisne spremenljivke. Ta mera nam pove, koliko se je posamezen algoritem dejansko naučil. Če ima normalizirano standardno napako večjo od 1, je popolnoma neuporaben. Normalizirana standardna napaka je definirana kot

$$NS_\epsilon = \frac{S_\epsilon}{\hat{S}_\epsilon} \tag{3.2}$$

Pri tem je $\hat{S}_\epsilon$ standardna napaka metode, ki vedno napove povprečno vrednost.

3.2.3 LOSS

Za ocenjevanje kvantilnih napovedi je uporabljena funkcija $LOSS$ [15], ki meri obtežene absolutne odklone med primeri in kvantili

$$L_\alpha(y, q) = \begin{cases} \alpha|y - q| & y > q \\ (1 - \alpha)|y - q| & y \leq q \end{cases} \tag{3.3}$$

Pri tem je y prava vrednost, q pa napovedana vrednost za kvantil α.

3.2.4 PVNI, RPNI in RPNI-PVNI

Kvaliteto intervalnih napovedi smo merili z merami *PVNI*, *RPNI* in *RPNI-PVNI* iz [18]. *PVNI* je pokrivna verjetnost napovednih intervalov (angleško *Prediction interval covarage probability*) in je definirana kot odstotek testnih primerov, za katere je prava vrednost odvisne spremenljivke zajeta znotraj napovednih intervalov. *PVNI* je torej kvantizacija pravilnosti napovednih intervalov.

RPNI oziroma relativni povprečni napovedni interval (angleško *Relative mean prediction interval*) pa opisuje povprečno širino napovednih intervalov normalizirano s privzetim napovednim intervalom. Manjši kot je *RPNI*, bolj optimalni so intervali preverjane metode.

Kombinacija obeh mer uspešnosti napovednih intervalov *RPNI-PVNI* omogoča lažjo primerjavo več metod preko različnih domen. Predstavljena pa je s formulo:

$$RPNI\text{-}PVNI = 100 \cdot RPNI + log\Big(max\Big((PVNI^* - PVNI)^2, 10^{-10}\Big)\Big), \quad (3.4)$$

kjer je PVNI* zastavljena pokrivna verjetnost napovednih intervalov. Logaritemski doprinos pa je omejen, saj se sicer približuje minus neskončno, ko se PVNI približuje zastavljeni vrednosti.

Poleg kvalitet napovedi pa smo preverjali tudi čas izvajanja posameznih metod. Merili smo potreben čas za učenje in napovedovanje preko vseh kvantilov za vsako množico podatkov posebej.

Vsa testiranja so bila opravljena s petkratnim prečnim preverjanjem (angleško *5-fold cross-validation*) [21].

Poglavje 4

Analiza rezultatov

V tem poglavju analiziramo rezultate primerjave metod z zgoraj opisanimi cenilkami. Vsi izračuni so bili narejeni v programskem okolju R [19], ki je namenjen statistični obdelavi podatkov in zelo priljubljen tudi na področju strojnega učenja. Vsebuje veliko algoritmov iz tega področja in tudi ogromno zbirko podatkovnih množic za modeliranje.

4.1 Standardna napaka

V tabeli 4.1 so izračunane standardne napake za posamezne metode in za vse umetne množice. V prvem stolpcu tabele so podatkovne množice, na desni strani pa so vrednosti standardnih napak. Pri tem smo dodatno s sivimi horizontalnimi črtami ločili različne vrste problemov v učnih množicah.

Na linearnem problemu se pričakovano najbolje izkaže metoda KR, ki ima standardno napako blizu 0 (precej manjšo od 10^{-3}). Metodi KRG in ENKG delujeta precej slabše. Njuni različici z uporabljenim samo enim atributom za točko delitve in privzeta ($K = \sqrt{n}$) sta najslabši. Različici, ki uporabljata vse neodvisne spremenljivke, pa delujeta precej bolje, kar je posledica majhnega števila atributov v podatkih, ki uspešnost metod KRG in ENKG precej omejijo.

Pri kvadratnih problemih se že pokažejo prednosti metod KRG in ENKG.

	KR	KRG^1	KRG	KRG^n	ENKG^1	ENKG	ENKG^n
LIN	0.00	3.79	3.14	2.95	3.74	2.73	2.65
SQR	25.2	43.1	35.5	35.8	45.8	33.1	30.8
SRN	28.7	30.8	22.8	22.2	32.9	22.4	19.5
SGN	33.1	26.2	18.5	16.2	32.2	18.8	14.2
SRV	12.1	16.0	10.1	8.13	19.6	9.82	7.00
SNS	6.19	6.15	6.07	6.25	6.25	5.87	6.00
SLN	12.7	12.7	12.5	12.7	12.9	12.6	13.8
CPX	49.5	32.8	19.5	14.8	45.3	19.2	12.3
RND	94.0	97.7	99.1	98.3	94.6	96.9	100

Tabela 4.1: Standardna napaka na umetnih podatkovnih množicah

Na brezšumni učni množici spet najbolje deluje metoda KR. Na podatkih z dodanim šumom pa sta metodi KRG in ENKG bistveno boljši. Različici, ki uporabljata samo eno neodvisno spremenljivko, imata še vedno nekaj težav. Različice s privzetim število atributov in z vsemi atributi pa se odrežejo precej boljše. Najbolj učinkoviti sta ravno različici z vsemi atributi, ob tem pa je najboljša metoda ENKG^n. Največ težav vsem metodam povzroči Gaussov šum, medtem ko odvečna spremenljivka nima tako velikega vpliva. Pri naključnem šumu pa so najmanjše razlike med vsemi metodami.

Majhne razlike med točnostjo metod so tudi pri sinusnih problemih. Tu sta najboljši privzeti različici metod KRG in ENKG. Ostale različice in tudi metoda KR pa so jima zelo blizu. Linearni šum pa ne povzroči večjih težav nobeni metodi in tudi ne naredi večje razlike med njihovo uspešnostjo.

Pri kompleksnejšem problemu ima pričakovano največ težav metoda KR. Najbolje se odreže ENKG^n, ki ji sledi KRG^n, privzeti različici KRG in ENKG sta v sredini, različici z uporabljenim samo enim atributom za točko rezanja pa sta precej za njimi.

Pri podatkih z naključno odvisno spremenljivko pričakovano vse metode delujejo precej slabo, med njimi pa ni večjih razlik v uspešnosti. Zanimivo je,

da ima najmanjšo standardno napako metoda KR. Glede na naravo te podatkovne množice pa ne moremo delati kakšnih posebnih sklepov o učinkovitosti metod.

	KR	KRG^1	KRG	KRG^n	ENKG^1	ENKG	ENKG^n
airquality	21.0	19.0	18.6	18.4	19.2	18.0	18.4
Boston	6.24	5.67	5.30	5.51	5.34	5.28	5.22
NO2	0.55	0.50	0.49	0.50	0.52	0.50	0.50
ozone	4.56	3.91	4.06	4.17	3.95	3.91	4.08
swiss	7.73	8.31	7.68	7.30	8.61	8.06	8.30

Tabela 4.2: Standardna napaka na realnih podatkovnih množicah

Na realnih problemih v tabeli 4.2 vidimo, da se v skoraj vseh primerih najslabše odreže KR. Pri vseh različicah metod KRG in ENKG pa ni večjih razlik. Najslabši sta različici z uporabljenim enim atributom za točko delitve. Privzeti različici in različici z vsemi neodvisnimi spremenljivkami pa imata precej podobno napako. V povprečju je še najbolj učinkovita metoda ENKG, a s skoraj nično prednostjo pred KRG. Ob tem lahko sklepamo, da večje število atributov torej izniči večje razlike med različicami metod KRG in ENKG.

4.2 Normalizirana standardna napaka

Metoda KR ponovno izstopa na linearnem problemu, kjer ima tudi normalizirano standardno napako blizu 0 (precej manjšo od 10^{-3}). Sicer pa v tabeli 4.3 vidimo, da se, razen pri podatkih z naključno generirano odvisno spremenljivko (RND), vse metode precej naučijo. Pri podatkih RND pa to niti ni mogoče. Ob tej izjemi so vse metode še najmanj uspešne na sinusnih problemih.

Tudi na realnih podatkih (tabela 4.4) se izkaže, da so se vse metode nekaj naučile. Njihova uspešnost je po normalizirani standardni napaki nekoliko

	KR	KRG^1	KRG	KRG^n	ENKG^1	ENKG	ENKG^n
LIN	0.000	0.412	0.341	0.320	0.406	0.297	0.288
SQR	0.248	0.424	0.350	0.353	0.451	0.326	0.303
SRN	0.408	0.438	0.324	0.316	0.468	0.319	0.277
SGN	0.458	0.363	0.256	0.224	0.446	0.260	0.197
SRV	0.296	0.392	0.247	0.199	0.480	0.240	0.171
SNS	0.746	0.741	0.732	0.753	0.753	0.708	0.723
SLN	0.798	0.798	0.785	0.798	0.811	0.792	0.867
CPX	0.465	0.308	0.183	0.139	0.425	0.180	0.116
RND	1.150	1.200	1.220	1.210	1.160	1.190	1.230

Tabela 4.3: Normalizirana standardna napaka na umetnih podatkih

manjša kot pri umetnih podatkih. So se pa metode največ naučile na podatkih *ozone*, najmanj pa na podatkih *Boston* in *NO2*.

	KR	KRG^1	KRG	KRG^n	ENKG^1	ENKG	ENKG^n
airquality	0.635	0.574	0.562	0.556	0.580	0.544	0.556
Boston	0.747	0.678	0.634	0.659	0.639	0.632	0.625
NO2	0.732	0.666	0.652	0.666	0.692	0.666	0.666
ozone	0.570	0.489	0.508	0.521	0.494	0.489	0.510
swiss	0.605	0.650	0.601	0.571	0.674	0.631	0.650

Tabela 4.4: Normalizirana standardna napaka na realnih podatkih

4.3 Napaka LOSS

Uspešnost metod, ocenjenih z napako LOSS, je prikazana z Demšarjevimi diagrami [5] za prikaz primerjave več metod, testiranih na več različnih podatkovnih množicah. Zgoraj je os, na kateri prikažemo povprečne range metod tako, da so na levi strani najmanjši (najboljši) rangi in v desno smer večji

(slabši) rangi. Pri primerjavi več algoritmov med sabo so skupine algoritmov, ki niso značilno različni, povezane z odebeljeno vodoravno črto. Nad grafom pa je prikazana tudi kritična razlika (CD; angleško *Critical difference*), ki pove, kdaj je kateri od algoritmov značilno drugačen od ostalih.

Povprečni rangi so izračunani s Friedmanovim testom [8, 9]. Ta je neparametričen in deluje tako, da rangira uspešnost na vseh podatkovnih množicah za vsako metodo posebej in nato izračuna povprečje za posamezno metodo. Dobljene vrednosti povedo, katero mesto zaseda posamezna metoda med vsemi in so tako že informativne same po sebi.

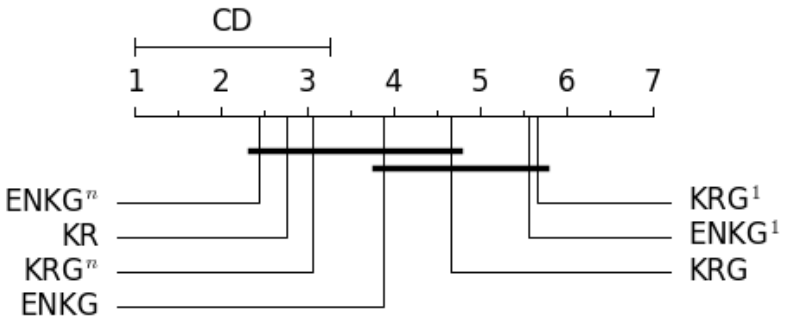
S Friedmanovim testom skušamo zavrniti ničelno hipotezo, da so vse metode enakovredne in bi morali biti njihovi rangi enaki. Če je ničelna hipoteza zavrnjena, je potrebno narediti še post-hoc test. V našem primeru smo delali Nemenyi test [17] za izračun kritične razlike pri $\alpha = 0.05$, ki je definirana kot

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{4.1}$$

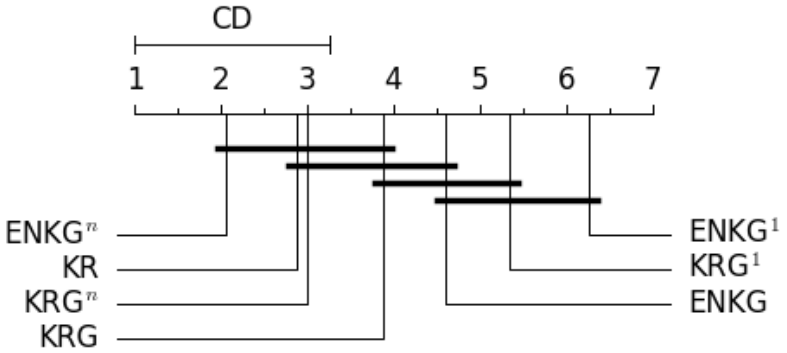
Pri tem je q_α kritična vrednost, k število algoritmov in N število podatkovnih množic, na katerih so bili preverjani algoritmi.

Na sliki 4.1 so zgoraj opisani grafi za prikaz uspešnosti metod po napaki LOSS. Pri tem so paroma združeni podatki za kvantile 0.01 in 0.99 (a), 0.05 in 0.95 (b), 0.25 in 0.75 (c), saj gre za intervale. Kvantil 0.5 (d) pa je srednja vrednost in tako nima svojega para.

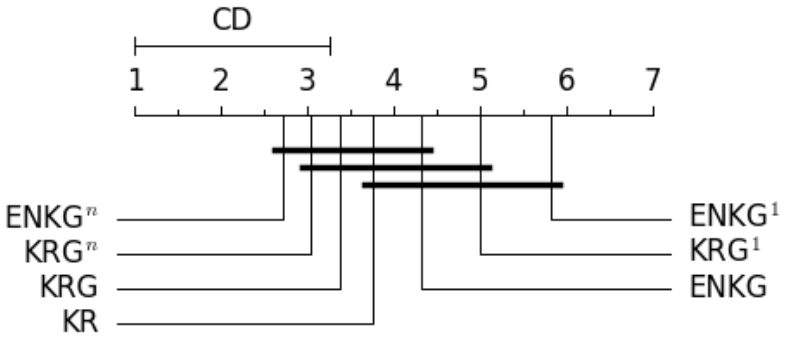
Pri napovedih za kvantila 0.01 in 0.99 (a) so le metode ENKG^n, KR in KRG^n značilno boljše od ostalih, med njimi pa ni značilnih razlik. Za presenetljivo dobro se je izkazala kvantilna regresija, ki je na drugem mestu po povprečnem rangu. Metodi ENKG in KRG s privzetimi nastavitvami sta v sredini, različici z uporabljenim samo enim atributom pa najslabši.



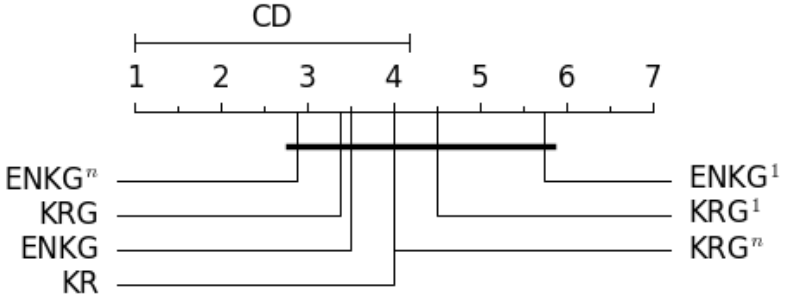
a) Primerjava uspešnosti pri napovedovanju za $q = 0.01$ in 0.99



b) Primerjava uspešnosti pri napovedovanju za $q = 0.05$ in 0.95



c) Primerjava uspešnosti pri napovedovanju za $q = 0.25$ in 0.75

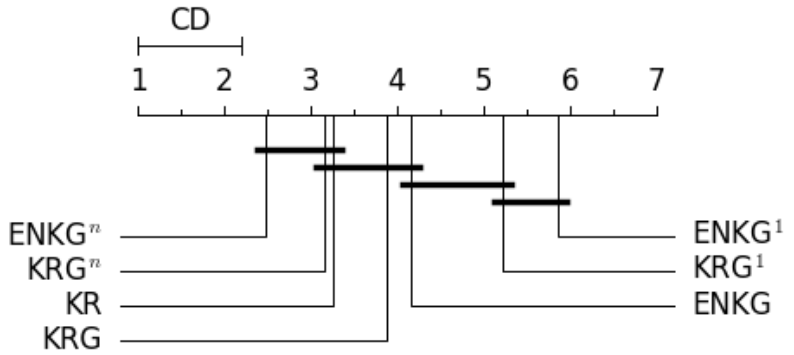


d) Primerjava uspešnosti pri napovedovanju za $q = 0.5$

Slika 4.1: Primerjava metod na umetnih podatkih za različne kvantile

Podobno je tudi pri napovedih za kvantila 0.05 in 0.95 (b), le da je več značilnih razlik med metodami. Uspešnost metod pa je po povprečnih rangih skoraj enaka. Prav tako je več značilnih razlik pri napovedih za kvantila 0.25 in 0.75 (c). Tudi uspešnost metod je precej podobna, le da je tu metoda KR manj uspešna in ni značilno boljša od nobene druge metode.

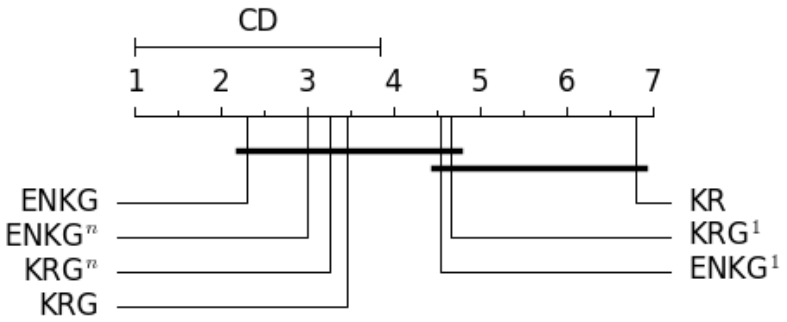
Tudi pri napovedih za srednjo vrednost (d) je vrstni red metod po uspešnosti precej podoben, vendar tu ni nobenih značilnih razlik med katerokoli metodo.



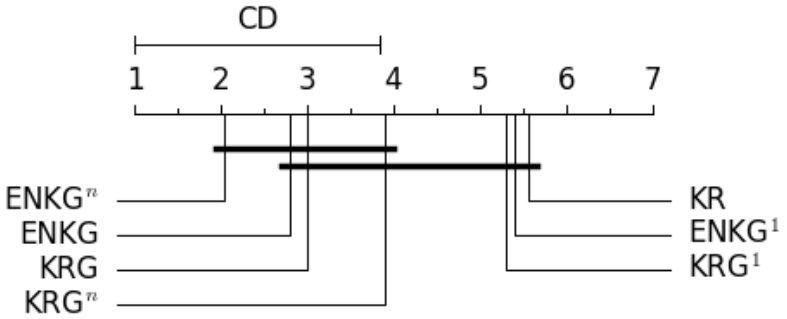
Slika 4.2: Primerjava metod na umetnih podatkih za napovedi vseh kvantilov

Za ugotavljanje, katera metoda je v splošnem najboljša, je na sliki 4.2 diagram uspešnosti metod na umetnih podatkih za napovedovanje vseh kvantilov. Najboljša je metoda ENKG^n, ki ni značilno boljša le od metod KRG^n in KR. Sledijo ji metode KRG^n, KR in KRG, ki so značilno boljše od najslabših treh, kjer sta kot najslabši različici KRG in ENKG, ki uporabljata samo eno neodvisno spremenljivko za napovedovanje.

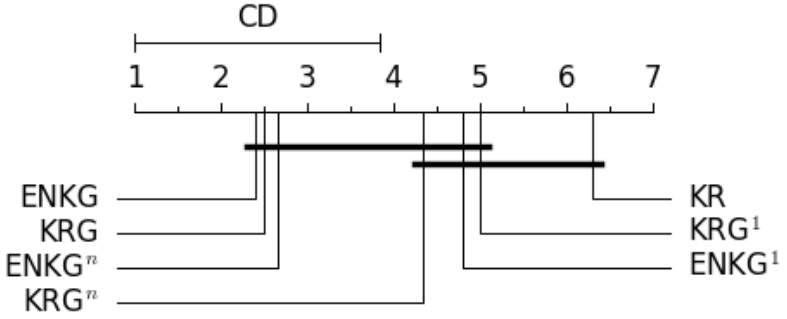
Pri realnih podatkih na sliki 4.3 so rezultati precej drugačni kot na umetno generiranih. Za napovedovanje kvantilov 0.01 in 0.99 (a) se najbolje izkažeta metodi ENKG in KRG, ter njuni različici, ki uporabljata vse atribute. Te metode so tudi značilno boljše od kvantilne regresije, ki velja za najslabšo.



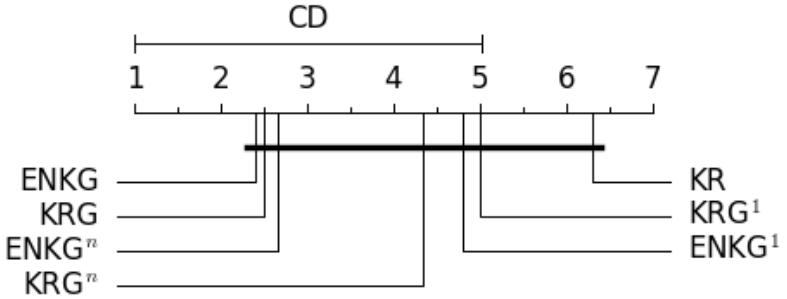
a) Primerjava uspešnosti pri napovedovanju za $q = 0.01$ in 0.99



b) Primerjava uspešnosti pri napovedovanju za $q = 0.05$ in 0.95



c) Primerjava uspešnosti pri napovedovanju za $q = 0.25$ in 0.75



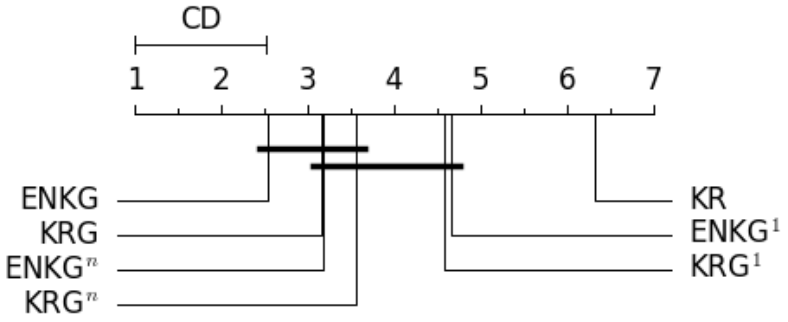
d) Primerjava uspešnosti pri napovedovanju za $q = 0.5$

Slika 4.3: Primerjava metod na realnih podatkih za različne kvantile

Za napovedi kvantilov 0.05 in 0.95 (b) sta ponovno najboljši metodi ENKG in KRG z različicama, ki uporabljata vse atribute. Pri tem je ENKG^n značilno boljša od najslabših treh metod. Med ostalimi pa ni značilnih razlik.

Pri napovedih za $q = 0.25$ in 0.75 (c) je podobno kot pri a), le da sta metodi ENKG in KRG s privzetimi parametri najboljši.

Značilnih razlik med uspešnostjo metod pa tako kot pri umetnih podatkovnih množicah tudi tu ni pri napovedih za srednjo vrednost (d).



Slika 4.4: Primerjava metod na realnih podatkih za napovedi vseh kvantilov

V splošnem, za napovedovanje vseh kvantilov, (slika 4.4), je najboljša metoda ENKG, ki je značilno boljša od najslabših treh (KR, ENKG^1 in KRG^1). Sledita ji metodi KRG in ENKG^n, ki nista značilno slabši, a sta značilno boljši le od kvantilne regresije, ki je tudi značilno najslabša med vsemi metodami. Tu se tudi vidi prednost ostalih metod na realnih podatkih, ki so običajno kompleksnejši. Prav tako je razvidno, da metode s privzetimi nastavitvami parametrov delujejo bolje.

4.4 Pokrivna verjetnost napovednih intervalov

Pri preverjanju uspešnosti intervalnih cenilk smo iz primerjave izločili različici KRG in ENKG, ki uporabljata samo eno neodvisno spremenljivko, saj

sta ti v skoraj vseh primerih daleč najslabši (med različicami metod KRG in ENKG). So pa v primerjavo vključene metode najbližjih sosedov, stremljenja in maksimalne verjetnosti ter njuna kombinacija.

V tabeli 4.5 so vrednosti PVNI na umetnih podatkovnih množicah. Na linearnem in kvadratnih problemih je najboljša metoda KRG^n, ki se najbolj približa zastavljeni PVNI (0.95). Na sinusnih problemih sta najboljši metodi NS in SMV. Na kompleksnem pa se za najboljšo izkaže metoda ENKG^n. V splošnem pa ima, z izjemo pri linearnih podatkih, najmanjše pokrivne verjetnosti napovednih intervalov metoda KR. Največje in hkrati tudi prevelike vrednosti PVNI pa dajejo metode, ki ne uporabljajo kvantilov (NS, SMV in NS-SMV). Kot najboljši se tako izkažeta metodi KRG in ENKG ter njuni različici z vsemi atributi.

	KR	KRG	KRG^n	ENKG	ENKG^n	NS	SMV	NS-SMV
LIN	0.995	0.975	0.945	0.980	0.970	0.990	0.995	0.990
SQR	0.855	0.975	0.960	0.970	0.970	0.980	0.990	0.990
SRN	0.875	0.970	0.935	0.975	0.980	0.980	0.985	0.980
SGN	0.865	0.965	0.920	0.975	0.945	0.975	0.980	0.975
SRV	0.870	0.990	0.945	0.990	0.965	0.985	0.995	0.990
SNS	0.870	0.950	0.905	0.940	0.900	0.965	0.970	0.970
SLN	0.880	0.910	0.870	0.895	0.870	0.970	0.960	0.975
CPX	0.860	0.985	0.900	0.980	0.935	0.990	0.990	0.990

Tabela 4.5: PVNI na umetnih podatkovnih množicah

Tudi na realnih podatkih (tabela 4.6) največje vrednosti PVNI dajejo metode, ki ne uporabljajo kvantilov. So pa tu precej bližje zastavljeni vrednosti PVNI in so med najboljšimi, ker imajo ostale metode precej manjše PVNI. Najbližje sta jima metodi KRG in ENKG, ki sta pri podatkih *airquality* celo najboljši. Najslabša pa je ponovno metoda KR z najmanjšimi vrednostmi PVNI.

	KR	KRG	KRG^n	ENKG	ENKG^n	NS	SMV	NS-SMV
airquality	0.847	0.945	0.855	0.927	0.901	0.982	0.973	0.973
Boston	0.868	0.913	0.879	0.869	0.822	0.974	0.968	0.968
NO2	0.888	0.898	0.862	0.916	0.856	0.958	0.956	0.956
ozone	0.864	0.936	0.927	0.936	0.918	0.964	0.955	0.961
swiss	0.658	0.896	0.787	0.873	0.744	0.956	0.978	0.956

Tabela 4.6: PVNI na realnih podatkovnih množicah

4.5 Relativni povprečni napovedni interval

	KR	KRG	KRG^n	ENKG	ENKG^n	NS	SMV	NS-SMV
LIN	0.000	0.626	0.466	0.670	0.462	0.530	0.659	0.594
SQR	0.252	0.634	0.498	0.686	0.502	0.561	0.672	0.617
SRN	0.377	0.630	0.478	0.701	0.476	0.620	0.714	0.667
SGN	0.370	0.479	0.267	0.586	0.266	0.589	0.689	0.639
SRV	0.277	0.520	0.288	0.586	0.277	0.555	0.659	0.607
SNS	0.727	0.794	0.710	0.813	0.697	0.944	0.959	0.952
SLN	0.721	0.819	0.771	0.832	0.728	1.016	1.024	1.020
CPX	0.404	0.441	0.153	0.586	0.152	0.633	0.752	0.693

Tabela 4.7: RPNI na umetnih podatkovnih množicah

Metoda KR tako dobro deluje na linearnih problemih, da je relativni povprečni napovedni interval (RPNI) bistveno manjši od 10^{-3} in tako v tabeli 4.7 zaokrožen na 0. Tudi na ostalih podatkovnih množicah daje metoda KR skoraj najmanjše napovedne intervale, vendar pa smo zgoraj videli, da tam nima tako dobre pokrivne verjetnosti napovednih intervalov. Majhne intervale dajeta tudi metodi ENKG^n in KRG^n. Njuni različici s privzetimi nastavitvami nekoliko zaostajata in sta primerljivi z metodo NS. Največje relativne napovedne intervale pa daje metoda SMV. Kombinirana metoda NS-SMV pa ima, pričakovano, vrednosti RPNI približno v povprečju me-

tod NS in SMV. Zanimivo je, da se največji relativni napovedni intervali pojavljajo pri sinusnih problemih in ne pri kompleksnem, kot bi pričakovali.

	KR	KRG	KRG^n	ENKG	ENKG^n	NS	SMV	NS-SMV
airquality	0.639	0.581	0.490	0.641	0.519	0.853	0.884	0.868
Boston	0.632	0.418	0.401	0.421	0.394	0.774	0.755	0.765
NO2	0.765	0.769	0.700	0.784	0.680	0.974	0.954	0.964
ozone	0.534	0.542	0.525	0.526	0.479	0.648	0.613	0.630
swiss	0.367	0.605	0.542	0.545	0.468	0.822	0.721	0.771

Tabela 4.8: RPNI na realnih podatkovnih množicah

Na realnih podatkih (tabela 4.8) so razlike med metodami manjše. Za najboljšo po meri RPNI se izkaže ENKG^n, ki je zgolj pri podatkih *airquality* nekoliko slabša od druge najboljše metode, KRG^n. Sledijo jima metode KRG in ENKG s privzetimi nastavitvami ter metoda KR. Metode, ki ne uporabljajo kvantilov, pa so precej slabše. Ponovno pa je najslabša metoda SMV.

4.6 Kombinirana statistika RPNI-PVNI

Kombinirana statistika s kombinacijo metod RPNI in PVNI daje vpogled v to, katera metoda je najbolj optimalna. Pri umetnih podatkovnih množicah (tabela 4.9) je za linearni problem daleč najbolj optimalna metoda KR. Zaradi zelo majhnih relativnih povprečnih napovednih intervalov in pokrivnih verjetnosti napovednih intervalov, ki se zelo približujejo zastavljeni pokrivni verjetnosti, ima RPNI-PVNI vrednost celo negativno. Med najboljšimi je tudi na kvadratnih problemih, kjer je samo pri podatkih *SGN* od nje boljša metoda KRG^n. V ostalih primerih sta najboljši metodi ENKG^n in KRG^n, ki imata zelo podobne vrednosti RPNI-PVNI. Njuni različici s privzetimi nastavitvami nekoliko zaostajata in sta podobno dobri kot metoda NS. Najslabša je metoda SMV. Kombinirana metoda NS-SMV pa ima tudi pri statistiki

RPNI-PVNI vrednosti v povprečju metod NS in SMV.

	KR	KRG	KRG^n	ENKG	ENKG^n	NS	SMV	NS-SMV
LIN	−3.17	59.4	43.4	63.9	43.0	49.8	62.7	56.2
SQR	20.5	58.7	45.0	63.9	45.5	51.4	62.5	57.0
SRN	32.5	57.8	42.6	64.9	42.4	56.8	66.2	61.5
SGN	32.0	43.0	21.7	53.7	21.7	54.0	64.0	59.0
SRV	22.7	47.0	23.7	53.6	22.7	50.4	60.9	55.6
SNS	67.6	74.3	66.0	76.3	64.7	89.4	90.9	90.1
SLN	67.0	76.9	72.0	78.2	67.8	96.6	97.4	97.0
CPX	35.6	39.3	10.5	53.8	10.4	58.5	70.4	64.5

Tabela 4.9: RPNI-PVNI na umetnih podatkovnih množicah

Na realnih podatkih (tabela 4.10) sta najbolj optimalni metodi KRG^n in ENKG^n. Slednja je v manjši prednosti, z izjemo pri podatkih *swiss*, kjer je daleč najboljša metoda KR. Sicer pa jima sledita metodi KRG in ENKG s precej podobnimi vrednostmi RPNI-PVNI kot jih ima tudi metoda KR. Najslabše pa so metode, ki ne uporabljajo kvantilov (NS, SMV, NS-SMV).

	KR	KRG	KRG^n	ENKG	ENKG^n	NS	SMV	NS-SMV
airquality	59.4	53.5	44.4	59.5	47.3	80.7	83.9	82.3
Boston	59.1	37.7	36.0	38.0	35.2	73.2	71.4	72.3
NO2	71.8	72.2	65.3	73.7	63.2	92.7	90.7	91.7
ozone	48.5	49.3	47.6	47.7	43.0	59.9	56.4	58.1
swiss	34.3	58.0	51.7	52.0	44.4	79.7	69.6	74.7

Tabela 4.10: RPNI-PVNI na realnih podatkovnih množicah

4.7 Čas izvajanja

Časi izvajanja na umetnih podatkovnih množicah (tabela 4.11) so pri vsaki metodi zelo podobni za vse podatke. Kompleksnost problema torej ne vpliva bistveno na čas izvajanja. Dimenzionalnost problemov je namreč enaka pri vseh množicah. Najhitrejša metoda je KR, sledita ji metodi ENKG in ENKG^n, ki je kljub uporabi vseh neodvisnih spremenljivk še vedno hitrejša od KRG in KRG^n. Slednji sta precej počasnejši zaradi permutacije podatkov, ki jih izvajata v fazi učenja. Obe metodi, ENKG in KRG pa približno enako upočasni število atributov, ki jih uporabljata njuni različici KRG^n in ENKG^n. Daleč najpočasnejše so metode, ki ne uporabljajo kvantilov. Ob tem je metoda NS nekoliko hitrejša od SMV, kombinirana metoda NS-SMV pa je še bistveno počasnejša. Ob tem naj omenimo, da slednja izračunava napovedne intervale metod NS in SMV, ter mora nato še izračunati povprečni meji intervalov. Njen čas izvajanja je tako nekoliko večji od vsote časov izvajanja metod NS in SMV.

	KR	KRG	KRG^n	ENKG	ENKG^n	NS	SMV	NS-SMV
LIN	0.036	0.416	0.566	0.206	0.330	1.586	2.406	3.930
SQR	0.032	0.376	0.542	0.164	0.302	1.534	2.300	4.216
SRN	0.028	0.392	0.550	0.290	0.244	1.586	2.350	3.976
SGN	0.024	0.384	0.546	0.208	0.306	1.590	2.230	3.994
SRV	0.028	0.394	0.532	0.216	0.296	1.556	2.332	3.774
SNS	0.028	0.392	0.694	0.214	0.378	1.550	2.244	3.680
SLN	0.038	0.418	0.610	0.260	0.366	1.762	2.600	4.212
CPX	0.036	0.392	0.658	0.216	0.344	1.546	2.254	3.722
SUM	0.25	3.16	4.70	1.77	2.57	12.7	18.7	31.5

Tabela 4.11: Čas izvajanja na umetnih podatkovnih množicah v sekundah

Tudi na realnih podatkih (tabela 4.12) je najhitrejša metoda KR, na katero tudi različne dimenzionalnosti problemov nimajo bistvenega vpliva.

	KR	KRG	KRG^n	ENKG	ENKG^n	NS	SMV	NS-SMV
airquality	0.030	0.226	0.324	0.154	0.168	0.960	1.156	1.966
Boston	0.062	1.964	3.686	0.840	1.392	9.942	12.850	20.592
NO2	0.040	1.164	1.814	0.588	0.762	7.218	10.088	17.144
ozone	0.038	0.800	1.390	0.378	0.762	3.762	5.044	8.572
swiss	0.044	0.124	0.138	0.050	0.114	0.506	0.672	0.968
SUM	0.21	4.28	7.35	2.01	3.20	22.4	29.8	49.2

Tabela 4.12: Čas izvajanja na realnih podatkovnih množicah v sekundah

Sledita ji metodi ENKG in ENKG^n, ki pa se jima precej pozna vpliv večdimenzionalnosti problema. Pri podatkih *Boston*, ki imajo največ primerov in neodvisnih spremenljivk, tako porabita bistveno več časa kot na ostalih podatkih z manjšo dimenzionalnostjo. Še precej počasnejši sta metodi KRG in KRG^n. Daleč najpočasnejše pa so metode, ki ne uporabljajo kvantilov. Tudi na slednjih se pozna vpliv večdimenzionalnosti problemov in so ravno tako najpočasnejše pri podatkih *Boston*.

4.8 Vpliv števila dreves

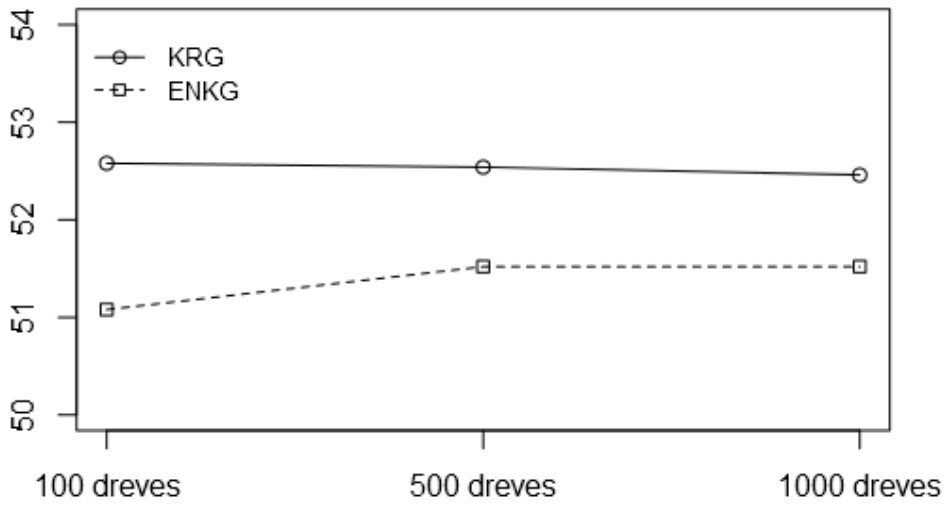
	KRG^{100}	KRG^{500}	KRG^{1000}	ENKG^{100}	ENKG^{500}	ENKG^{1000}
airquality	53.2	51.8	52.2	55.5	55.8	55.2
Boston	40.5	40.4	40.1	35.0	36.4	36.2
NO2	70.2	70.3	70.8	72.7	72.7	72.8
ozone	44.3	45.0	44.8	42.6	42.7	43.1
swiss	54.7	55.2	54.4	49.6	50.0	50.3

Tabela 4.13: RPNI-PVNI na realnih podatkih z različnim številom dreves

Na realnih podatkih smo preverili tud vpliv števila dreves, ki jih gradita metodi KRG in ENKG (s privzetimi nastavitvami). Metode imajo nadpisano

število dreves, ki so jih gradile. KRG^{100} torej označuje metodo KRG, ki je gradila 100 dreves.

V tabeli 4.13 vidimo, da število dreves nima ravno veliko vpliva na uspešnost napovedovanja po statistiki RPNI-PVNI. Pri metodi KRG sta različici, ki gradita 100 in 500 dreves, boljši v treh primerih, v dveh pa sta slabši od različice, ki gradi 1000 dreves. Pri metodi ENKG pa je različica, ki gradi 100 dreves, boljša v vseh primerih od različice, ki gradi 500 dreves, in slabša le v enem primeru od različice, ki gradi 1000 dreves. Slednja je v dveh primerih boljša od različice, ki gradi pol manj dreves. Za privzeto vrednost je tako 100 dreves dobra izbira.



Slika 4.5: Povprečni RPNI-PVNI za metodi pri različnem številu dreves

Iz povprečnih vrednostih RPNI-PVNI na sliki 4.5 se vidi, da so razlike med različicami znotraj posamezne metode zelo majhne. Majhna je tudi razlika med različicami metode KRG in različicami metode ENKG.

Izmerili smo tudi čase, ki jih potrebujejo metode z različnim številom dreves za učenje in napovedovanje. V tabeli 4.14 vidimo, da različici, ki gradita 500 in 1000 dreves, potrebujeta bistveno več časa kot različica s 100

	KRG^{100}	KRG^{500}	KRG^{1000}	$ENKG^{100}$	$ENKG^{500}$	$ENKG^{1000}$
airquality	0.216	0.978	2.562	0.162	0.704	1.364
Boston	1.732	8.884	16.522	0.754	4.266	8.738
NO2	1.268	6.488	12.360	0.650	3.224	6.794
ozone	0.768	3.834	8.762	0.520	2.192	4.838
swiss	0.122	0.410	0.812	0.048	0.382	0.618
SUM	4.11	20.6	41.1	2.13	10.8	22.4

Tabela 4.14: Čas izvajanja metod KRG in ENKG z različnim številom dreves na realnih podatkih

drevesi, pri obeh metodah. Še posebej razvidno pa je to pri metodi KRG, ki je že v osnovi precej počasnejša od metode ENKG.

4.9 Vpliv velikosti učne množice

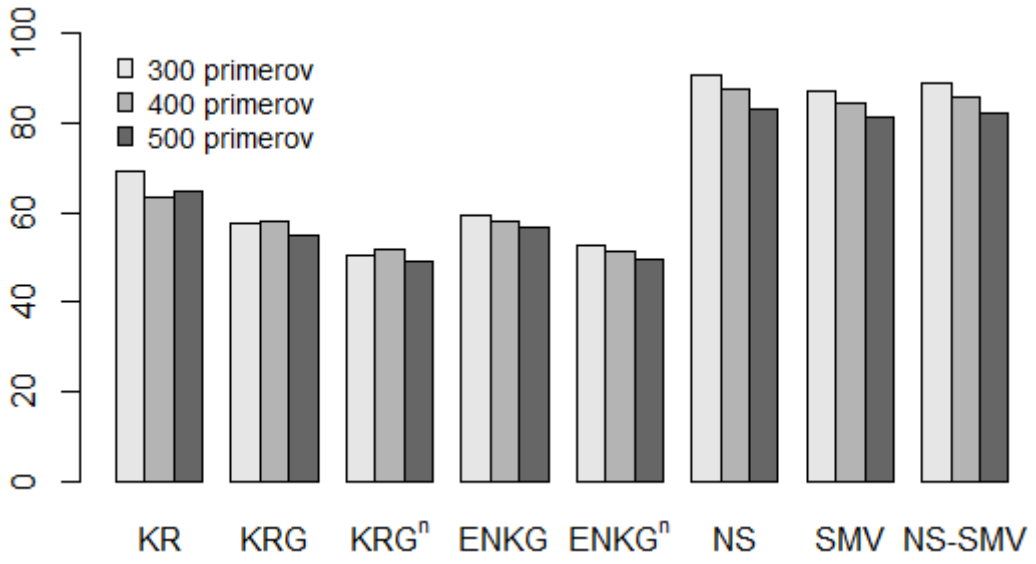
	KR	KRG	KRG^n	ENKG	$ENKG^n$	NS	SMV	NS-SMV
$Boston^{300}$	66.8	41.1	34.5	42.1	39.7	84.1	79.5	81.8
$Boston^{400}$	56.3	43.3	38.8	41.2	37.6	81.3	77.0	79.1
$Boston^{500}$	58.0	37.8	33.7	39.3	35.5	73.9	72.6	73.3
$NO2^{300}$	71.3	73.7	66.0	76.3	65.3	97.6	94.7	96.2
$NO2^{400}$	70.4	73.1	64.9	74.5	65.2	93.9	91.6	92.7
$NO2^{500}$	71.7	72.2	64.4	74.0	63.6	92.7	90.1	91.4

Tabela 4.15: RPNI-PVNI metod na realnih podatkih z različnim številom primerov

Vpliv velikosti učne množice na uspešnost metod s statistiko RPNI-PVNI smo preverili tako, da smo množici z realnimi podatki *Boston* in *NO2* omejili na 300, 400 in 500 primerov. Število primerov je v tabeli 4.15 nadpisano pri imenu podatkov.

Pri podatkih Boston vidimo, da so vse metode z izjemo KR najbolj uspešne pri največjem številu primerov v podatkovni množici. Ob tem pa imajo metode KRG in ENKG ter njuni različici z vsemi atributi najmanjše razlike v uspešnosti med množico z najmanjšim in največjim številom primerov. Zanimivo je, da sta metodi KRG in KRG^n pri množici s 300 primeri celo bolj uspešni kot pri množici s 400 primeri.

Manjše razlike v uspešnosti metod pa so na podatkih *NO2*. Razlog za to najverjetneje tiči v tem, da ta množica vsebuje zgolj pet neodvisnih spremenljivk, med tem ko jih je pri podatkih *Boston* kar 13. Ob tem metoda KR najbolje deluje na srednje veliki množici (400 primerov) in najslabše na največji (500 primerov). Ostale metode pa so najbolj uspešne na največji podatkovni množici in najmanj uspešne na množici z najmanj primeri.



Slika 4.6: RPNI-PVNI povprečen preko obeh učnih množic za različno število primerov

Izračunali smo tudi povprečne vrednosti RPNI-PVNI preko obeh množic (slika 4.6). Vidimo, da so pri metodah, ki gradijo gozdove, razlike precej majhne. Metodi KRG in KRG^n sta celo bolj učinkoviti na 300 kot na 400

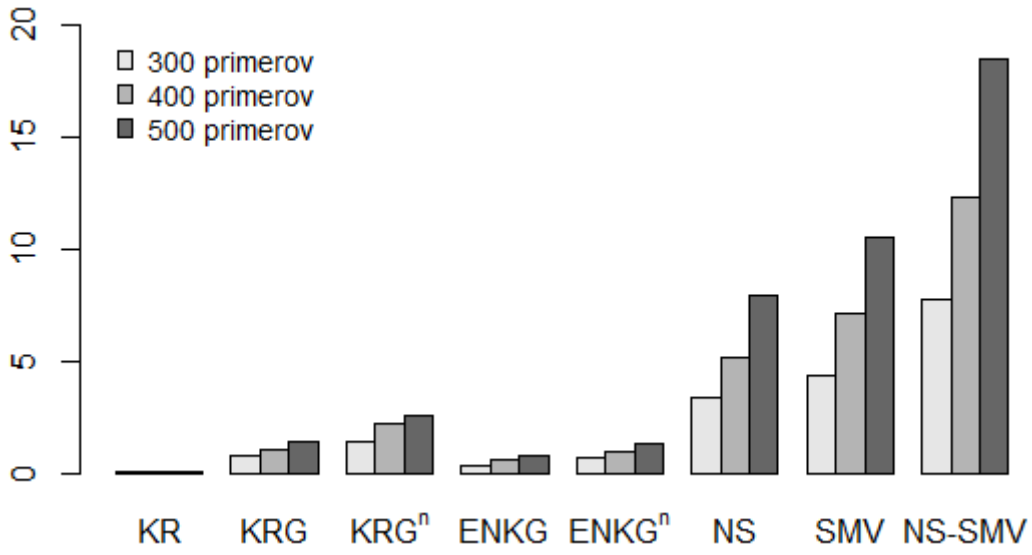
primerih. Metoda KR je najbolj učinkovita na 400 primerih. Večji vpliv števila primerov na RPNI-PVNI pa je opazen pri metodah, ki ne uporabljajo kvantilov.

	KR	KRG	KRG^n	ENKG	ENKG^n	NS	SMV	NS-SMV
Boston^{300}	0.086	0.920	1.840	0.424	0.820	4.178	5.094	9.040
Boston^{400}	0.066	1.220	2.976	0.846	1.188	5.868	8.028	13.660
Boston^{500}	0.062	1.570	3.304	0.778	1.956	8.712	11.926	20.556
NO2^{300}	0.032	0.610	0.930	0.294	0.482	2.490	3.578	6.432
NO2^{400}	0.034	0.878	1.378	0.400	0.704	4.358	6.228	10.900
NO2^{500}	0.038	1.172	1.786	0.714	0.724	7.062	9.160	16.454

Tabela 4.16: Čas izvajanja metod na realnih podatkih z različnim številom primerov

Razlike v časih izvajanja metod na različno velikih podatkovnih množicah (tabela 4.16) so le pri metodi KR zelo majhne v vseh primerih. Metode, ki ne uporabljajo kvantilov (NS, SMV, NS-SMV), so na najmanjši podatkovni množici vedno vsaj enkrat hitrejše kot na največji. Metoda ENKG in njena različica z vsemi atributi ima nekoliko manjšo razliko v izvajanju na najmanjši in največji podatkovni množici. Še nekoliko manjša razlika pa je pri metodama KRG in KRG^n. Vse metode z izjemo KR pa so precej hitrejše na podatkih *NO2* kot pri podatkih *Boston* (pri istem številu primerov). Slednje izhaja iz dejstva, da podatki *NO2* vsebujejo bistveno manj neodvisnih spremenljivk. Ob tem so razlike med izvajanjem na podatkih z več in manj atributi manjše na metodah, ki ne uporabljajo kvantilov.

Povprečni časi izvajanja vseh metod preko obeh podatkovnih množic so prikazani na sliki 4.7. Metoda KR ima tako majhne razlike, da so te neopazne. Pri ostalih metodah, ki uporabljajo kvantile, so te nekoliko bolj vidne. Pri metodah, ki ne uporabljajo kvantilov, pa so razlike zelo izrazite. Te metode so tudi občutno počasnejše od ostalih.



Slika 4.7: Čas izvajanja (v sekundah) povprečen preko obeh učnih množic za različno število primerov

4.10 Stabilnost metod

	KR	KRG	KRG^n	ENKG	ENKG^n	NS	SMV	NS-SMV
airquality	6.89	7.11	9.60	6.75	8.12	8.91	12.82	10.86
Boston	10.65	13.31	8.62	11.99	10.45	20.60	19.78	20.19
NO2	2.52	2.12	1.32	1.60	1.20	2.69	2.90	2.76
ozone	2.20	3.06	2.53	3.09	3.27	5.23	4.03	4.58
swiss	7.58	5.93	7.06	2.50	4.99	6.78	6.72	6.37
SUM	29.8	31.5	29.1	25.9	28.0	44.2	46.3	44.8

Tabela 4.17: Standardni odklon statistike RPNI-PVNI pri petkratnem prečnem preverjanju na realnih podatkih

Stabilnost metod smo preverili s standardnim odklonom izračunanih statistik RPNI-PVNI pri petkratnem prečnem preverjanju na realnih podatkih. V tabeli 4.17 vidimo, da se za najbolj stabilno metodo po uspešnosti z mero

RPNI-PVNI izkaže metoda ENKG. Nekoliko slabše so ostale metode, ki uporabljajo kvantile. Precej manj stabilne pa so metode, ki ne uporabljajo kvantilov. Ob tem se največja nestabilnost metod pokaže pri večdimenzionalnih podatkih *Boston*. Pri podatkih *NO2*, ki imajo skoraj enako primerov, a precej manj atributov kot jih je v podatkih *Boston*, pa vse metode delujejo precej stabilno.

Poglavje 5

Zaključek

Naredili smo pregled treh metod (KR, KRG in ENKG), ki za napovedovanje uporabljajo kvantile. Njihovo delovanje smo preverili na umetnih in realnih podatkovnih množicah, ter primerjali s tremi metodami, ki ne uporabljajo kvantilov (NS, SMV in NS-SMV). Analizirali smo vpliv različnih parametrov, ki jih uporabljajo metode, primerjali čase izvajanja v različnih pogojih in preverili različne vplive učnih množic na delovanje metod.

Ekstremno naključni kvantilni gozdovi so se izkazali za zelo konkurenčno metodo med vsemi primerjanimi. Pogosto so bili med najbolj uspešnimi in tudi časovno najbolj učinkovitimi. Pri zelo preprostih problemih je sicer najbolj uporabna kvantilna regresija, ki je zelo hitra in tudi z dobro točnostjo napovedi. Kvantilni regresijski gozdovi so po uspešnosti zelo blizu ekstremno naključnim kvantilnim gozdovom in v nekaj primerih celo boljši, a so v vseh primerih počasnejši. Metodi ENKG in KRG, ki uporabljata zgolj eno neodvisno spremenljivko, sta v splošnem precej slabši in tako ne ravno uporabni. Različici, ki uporabljata vse neodvisne spremenljivke, pa sta v kar nekaj primerih celo boljši od različic s privzetimi nastavitvami. Uporabni sta predvsem v primerih, ko imajo podatki malo neodvisnih spremenljivk. V splošnem, pa se za najbolj optimalno izbiro izkaže privzeta nastavitev števila atributov. Podobno je tudi pri izbiri števila dreves za obe metodi. Privzeta vrednost 100 dreves je sicer v nekaj primerih malo slabša od različic s 500 in

1000 drevesi, vendar je bistveno hitrejša. Pri izbiri velikosti učne množice pa je za vse metode najbolje izbrati čim večje število primerov. Metode ob tem sicer delujejo nekoliko počasneje, a so precej bolj uspešne. To je še posebej opazno pri podatkovnih množicah z večjim številom neodvisnih spremenljivk.

Metode, ki ne uporabljajo kvantilov (NS, SMV in NS-SMV) dajejo sicer intervale z zelo dobro pokrivno verjetnostjo napovednih intervalov, vendar so ti nekoliko manj optimalni kot pri metodama KRG in ENKG. Poleg tega pa so te metode tudi precej počasnejše.

Ekstremno naključni kvantilni gozdovi so torej dobra izboljšava običajnih naključnih dreves za napovedovanje kvantilov. Z dodatno naključnostjo pogosto prinesejo večjo natančnost napovedi in tudi pohitritev v delovanju. Izkažejo pa se tudi za stabilno metodo. Njihova stabilnost je precej boljša predvsem v primerjavi z metodami NS, SMV in NS-SMV.

Literatura

[1] D. Aha, D. W. Kibler, M. K. Albert, Instance-based learning algorithms, *Machine Learning*, 6:37-66, 1991

[2] S. Ahn, J. A. Fessler, Standard errors of mean, variance, and standard deviation estimators, *https://web.eecs.umich.edu/~fessler/papers/files/tr/stderr.pdf*, [dostopano 17. marec 2016]

[3] L. Breiman, Bagging predictors, *Machine Learning*, 24:123–140, 1996

[4] L. Breiman, Random forests, *Machine Learning*, 45:5-32, 2001

[5] J. Demšar, Statistical Comparisons of Classifiers over Multiple Data Sets, *Journal of Machine Learning Research*, 7:1-30, 2006

[6] T. G. Dietterich, Ensemble Methods in Machine Learning, *Multiple Classifier Systems, Volume 1857 of the series Lecture Notes in Computer Science*, 1-15, 2000

[7] K. Dwyer, R. Holte, Decision Tree Instability and Active Learning, *Machine Learning: ECML 2007*, 128-193, 2007

[8] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, *Annals of Mathematical Statistics*, 11:86–92, 1940

[9] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association*, 32:675–701, 1937

[10] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, *Machine Learning*, 63:3-42, 2006

[11] C. Gu, Smoothing Spline ANOVA Models: R Package gss, *Journal of Statistical Software*, 58(5):1-25, 2014

[12] R. Koenker, K. F. Hallock, Quantile Regression, *Journal of Economic Perspectives*, 15:143-156, 2001

[13] I. Kononenko, M. Kukar, Machine Learning and Data Mining: Introduction to Principles and Algorithms, *Horwood Publishing Limited*, 2007

[14] J. G. MacKinnon, Bootstrap Methods in Econometrics, *The Economic Record*, 82:2-18, 2006

[15] N. Meinshausen, Quantile Regression Forests, *Journal of Machine Learning Research*, 7:983-999, 2006

[16] S. J. Miller, The method of least squares, *Providence R1*, 2012

[17] P. B. Nemenyi, Distribution-free multiple comparisons, *PhD thesis, Princeton University*, 1963

[18] D. Pevec, Ocenjevanje zanesljivosti posameznih napovedi pri nadzorovanem učenju, *Doktorska disertacija, Fakulteta za računalništvo in informatiko*, 2013

[19] R Core Team, R: A Language and Environment for Statistical Computing, *https://www.R-project.org*, [dostopano 19. marec 2016]

[20] R Core Team, The R datasets package, *https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html*, [dostopano 19. marec 2016]

[21] P. Refaeilzadeh, L. Tang, H. Liu, Cross-Validation, *Encyclopedia of Database Systems*, 532-538, 2009

[22] F. W. Scholz, Maximum Likelihood Estimation, *Encyclopedia of Statistical Sciences*, 2006

[23] S. Seo, A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets, *Master's Thesis*, 2006

[24] W. N. Venables, B. D. Ripley, Modern Applied Statistics with S. Fourth Edition, *http://www.stats.ox.ac.uk/pub/MASS4*, [dostopano 21. marec 2016]